A CNN-RNN Framework for Image Annotation from Visual Cues and Social Network Metadata

Tobia Tesan\textsuperscript{1}, Pasquale Coscia\textsuperscript{2} and Lamberto Ballan\textsuperscript{2}

\textsuperscript{1}Quantexa Ltd, London, UK
\textsuperscript{2}University of Padova, Department of Mathematics, Italy
Visual Intelligence and Machine Perception (VIMP) Group
Image annotation

- Process of labelling images using text or annotation tools.
- Some images might be hard to recognize without additional context.
- Weakly-annotated images may help to disambiguate the visual classification task.
Image annotation

- Metadata of images shared on social-media are an ideal source of additional information.

**DSCN9999233.JPG**

*Taken: May 30 2015, 9:15*

*Tags: 叶子, 树, 微距摄影*

**Exploring Tag: 叶子**
Metadata Limitations

- Image metadata are useful but can be:
  - noisy
  - highly subjective

- Models should also be robust to vocabulary changes.

vocabulary = [dog, cat, ... cute, laptop]

vocabulary = [dog, cat, ... cute, laptop, beard]
Our approach

- Advanced semantic mapping and CNN-RNN fusion schemes.
- Visual features and metadata to jointly leverage context and visual cues.
- State-of-the-art results on the multi-label image annotation task using the NUS-WIDE dataset.
- Our models decrease both sensory and semantic gaps to better annotate images.

Context (tags) + Visual Cues
Visual models vs Joint Models
Metadata Encoding

One-hot Encoding

\[ o_x = \sum_{i \text{ s.t. } t_i \in \{t_1, t_2, \ldots, t_n\}} e_i^x \]

Semantic-aware Encoding

Word2vec

WordNet

\[ \rho(o_x; \beta) = \sum_{i=1}^{\tau} o_{x(i)} \cdot \beta(t_{(i)}) \]
Visual Models

Visual only

LTN

RTN
Joint Models

LTN+Vecs

LTN+AllVecs

LTwin
Joint Models

LTN+Vecs

LTN+AllVecs

LTwin

LTwin+RNN

LTwin+2RNN

LZip
## Dataset & Metrics

- **NUS-WIDE dataset:**
  - 269,648 images collected from Flickr;
  - 81 labels (manual annotation);
  - 5000 most frequent tags.

- **Metrics:**
  - Per-label/per-image mean Average Precision (mAP);
  - Precision and recall.

<table>
<thead>
<tr>
<th>Image</th>
<th>Label</th>
<th>Metadata (tags)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image: 163792</td>
<td>grass</td>
<td>centipede, yellow, naturesfinest, k100d, macro, pentax, kit, eyes, animals, grass, chenille, nature, 1855, johannpix, caterpillar</td>
</tr>
<tr>
<td>Neighbour: 140470</td>
<td>animal</td>
<td>flickrdiamond, animalkingdomelite, dragonfly, naturesfinest, k100d, macro, pentax, wild, kit, animals, blue, damselfly, green, nature, blueribbonwinner, 1855, diamondclassphotographer, closeup, johannpix, libellule</td>
</tr>
<tr>
<td>Neighbour: 140175</td>
<td>sun, sky, flowers, clouds</td>
<td>yellow, naturesfinest, k100d, pentax, flash, see, kit, outdoors, overtheshot, 1855, colors, sun, flowers, johannpix, sky, tulips, fillin</td>
</tr>
<tr>
<td>Neighbour: 15106</td>
<td>animal</td>
<td>yellow, macro, 5hits, selectivecolorization, animals, selectivecolor, bird, nature, chicken, chick, beak, baby, bw</td>
</tr>
</tbody>
</table>
Experimental Results (1/4)

- Our best results in comparison to several baselines and SOTA models.

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP$_{lab}$</th>
<th>mAP$_{img}$</th>
<th>rec$_{lab}$</th>
<th>prec$_{lab}$</th>
<th>rec$_{img}$</th>
<th>prec$_{img}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tag-only Model + linear SVM [7]</td>
<td>46.67</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Graphical Model (all metadata) [7]</td>
<td>49.00</td>
<td>-</td>
<td>35.60</td>
<td>31.65</td>
<td>60.49</td>
<td>48.59</td>
</tr>
<tr>
<td>CNN + WARP [16]</td>
<td>-</td>
<td>-</td>
<td>30.40</td>
<td>40.50</td>
<td>61.70</td>
<td>49.90</td>
</tr>
<tr>
<td>CNN-RNN [21]</td>
<td>-</td>
<td>-</td>
<td>50.17 *</td>
<td>55.65 *</td>
<td>71.35 *</td>
<td>70.57 *</td>
</tr>
<tr>
<td>SR-RNN [22]</td>
<td>-</td>
<td>-</td>
<td>58.52 *</td>
<td>63.51 *</td>
<td>77.33 *</td>
<td>76.21 *</td>
</tr>
<tr>
<td>SR-RNN + Vecs [22] †</td>
<td>-</td>
<td>-</td>
<td>60.00</td>
<td>80.60</td>
<td>41.50 *</td>
<td>70.40 *</td>
</tr>
<tr>
<td>SRN [35]</td>
<td>60.00</td>
<td>80.60</td>
<td>58.70 *</td>
<td>81.10 *</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MangoNet [33]</td>
<td>62.80</td>
<td>80.80</td>
<td>59.90 *</td>
<td>80.60 *</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LTN [2]</td>
<td>52.78 ±0.34</td>
<td>80.34 ±0.07</td>
<td>43.61 ±0.47</td>
<td>46.98 ±1.01</td>
<td>74.72 ±0.16</td>
<td>53.69 ±0.13</td>
</tr>
<tr>
<td>LTN + Vecs [2] †</td>
<td>61.88 ±0.36</td>
<td>80.27 ±0.08</td>
<td>57.30 ±0.44</td>
<td>54.74 ±0.63</td>
<td>75.10 ±0.20</td>
<td>53.46 ±0.09</td>
</tr>
<tr>
<td>Upper bound</td>
<td>100.00 ±0.00</td>
<td>100.00 ±0.00</td>
<td>65.82 ±0.35</td>
<td>60.68 ±1.32</td>
<td>92.09 ±0.10</td>
<td>66.83 ±0.12</td>
</tr>
<tr>
<td>Our baseline: v-only</td>
<td>45.05 ±0.11</td>
<td>76.88 ±0.11</td>
<td>42.31 ±0.59</td>
<td>43.74 ±1.07</td>
<td>71.41 ±0.13</td>
<td>51.36 ±0.13</td>
</tr>
<tr>
<td>Our baseline: LTN$_{n:id}$</td>
<td>53.17 ±0.12</td>
<td>79.82 ±0.16</td>
<td>45.67 ±1.75</td>
<td>47.64 ±2.18</td>
<td>74.29 ±0.13</td>
<td>53.34 ±0.17</td>
</tr>
<tr>
<td>Our baseline: LTN + Vecs$_{n:id,f:id}$ †</td>
<td>54.86 ±0.20</td>
<td>81.34 ±0.15</td>
<td>46.56 ±1.39</td>
<td>50.10 ±1.70</td>
<td>75.67 ±0.17</td>
<td>54.37 ±0.14</td>
</tr>
<tr>
<td>Our model: RTN$_{n:w2v}$</td>
<td>55.36 ±0.34</td>
<td>79.77 ±0.27</td>
<td>48.73 ±2.77</td>
<td>51.21 ±2.61</td>
<td>74.35 ±0.29</td>
<td>53.28 ±0.24</td>
</tr>
<tr>
<td>Our model: LTwin$_{n:w2v,f:w2v}$ †</td>
<td>63.13 ±0.31</td>
<td>83.77 ±0.06</td>
<td>54.40 ±1.33</td>
<td>51.86 ±1.58</td>
<td>78.06 ±0.05</td>
<td>55.78 ±0.13</td>
</tr>
</tbody>
</table>
Experimental Results (2/4)

- $mAP_{lab}$ and $mAP_{img}$ for visual models.
Experimental Results (3/4)

- $mAP_{lab}$ and $mAP_{img}$ for joint models (word2vec embeddings).
Experimental Results (4/4)

- $mAP_{lab}$ and $mAP_{img}$ for joint models (wordNet embeddings).
Thank You!

3